BAYESIAN CLASSIFICATION OF PROTEOMICS BIOMARKERS FROM SELECTED REACTION MONITORING DATA USING AN ABC-MCMC APPROACH

A Thesis

by

KASHYAP NAGARAJA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Ullises Braga-Neto
Co-Chair of Committee,	Scott Miller
Committee Members,	Erchin Serpedin
	Alan R Dabney
Head of Department,	Miroslav Begovic

May 2019

Major Subject: Electrical Engineering

ABSTRACT*

Selective reaction monitoring (SRM) has become one of the main methods for low-mass range targeted proteomics by mass spectrometry (MS). However, in most SRM-MS biomarker validation studies the sample size is very small, and in particular smaller than the number of proteins measured in the experiment. Moreover, the data can be noisy due to a low number of ions detected per peptide by the instrument. In this paper, those issues are addressed by a model-based Bayesian method for classification of SRM-MS data, which relies on the SRM model proposed by Esmaeil[1] and collaborators and builds a kernel classifier, similarly to the classifier for LC-MS data proposed by Banerjee and Braga-Neto[3]. The methodology is likelihood-free, using Approximate Bayesian Computation (ABC) implemented via a Markov Chain Monte Carlo (MCMC) procedure and a kernel-based Optimal Bayesian Classifier (OBC). Extensive experimental results demonstrate that the proposed method is superior to classical methods, such as LDA and 3NN, when sample size is small, dimensionality is large, the data are noisy, or a combination of these.

*Reprinted with permission from *Bayesian Classification of Proteomics Biomarkers from Selected Reaction Monitoring Data using an Approximate Bayesian Computation-Markov Chain Monte Carlo Approach* by Kashyap Nagaraja and Ulisses Braga-Neto from Cancer Informatics journal Volume 17, pages 1:7, DOI:10.1177/1176935118786927, Copyright Date: May 24, 2018, Owners: Kashyap Nagaraja and Ulisses Braga-Neto

DEDICATION

To my parents.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude towards Dr. Ulisses Braga Neto for patiently guiding me throughout the thesis.I also would like to thank Mr. Upamanyu Banarjee for helping me inspite of his busy schedule.

I would like to thank my parents for their unwavering support as well as for teaching me the importance of perseverance and humility early on in life.

I would like to mention that the coursework at Texas A & M in general and the courses Pattern Recognition by Dr. Ulisses Braga Neto and Applied Multivariate Analysis by Dr. Alan Dabney helped me to understand the key concepts required for my thesis.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by Professor Ulisses Braga-Neto [advisor] of the ECEN Department at Texas A&M Unviersity.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was not supported via any funding sources.

NOMENCLATURE

ABC	Approximate Bayesian Computation
MCMC	Markov Chain Monte Carlo
LC-MS	Liquid chromatography-mass spectrometry
SRM-MS	Selected Reaction Monitoring -mass spectrometry
OBC	Optimal Bayesian Classifier
LDA	Linear Discriminant Analysis
KNN	K Nearest Neighbours

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION*

Proteomics is the field which deals with the study of cellular behavior and human disease at the protein level. Recently, cancer treatment and prevention have made great strides, thanks to the development of high-throughput technologies in proteomics. Among these, mass spectrometry (MS) analysis has become the preferred choice because of advantages such as high molecular specificity and better detection sensitivity[6]. Hence, MS is widely used in identification and quantification of complex proteome mixtures with the goal of discovering biomarkers, ie, molecular markers for disease[5].

However, a major challenge in biomarker discovery is the identification of low-abundance proteins in peripheral blood. Selected reaction monitoring (SRM), conducted using a triplequadrupole (QQQ) instrument, has an extended mass range and has become one of the main methods for low-mass-range targeted proteomics by MS[1].

Nevertheless, in most SRM-MS biomarker validation studies, the sample size is very small due to the economic cost of the experiments and difficulty in recruiting cases. Typically, the number of features (measured proteins) is vastly larger than the sample size. Moreover, depending on the instrument sensitivity, the data can be noisy due to low peptide efficiency, ie, low number of ions detected per peptide.

All the aforementioned issues create a difficult challenge to classical data-driven classification methods. In this article, this is addressed by a model-based Bayesian method for classification of SRM-MS data. We perform Bayesian inference of the parameters of the SRM model proposed in the work by Atashpaz-Gargari et al[7] and build a kernel classifier, similar to the classifier

*Reprinted with permission from *Bayesian Classification of Proteomics Biomarkers from Selected Reaction Monitoring Data using an Approximate Bayesian Computation-Markov Chain Monte Carlo Approach* by Kashyap Nagaraja and Ulisses Braga-Neto from Cancer Informatics journal Volume 17, pages 1:7, DOI:10.1177/1176935118786927, Copyright Date: May 24, 2018, Owners: Kashyap Nagaraja and Ulisses Braga-Neto

for liquid chromatography-mass spectrometry (LC-MS) data proposed in the work by Banerjee and BragaNeto[3]. As in the latter reference, our method uses a likelihoodfree approach, called approximate Bayesian computation (ABC)[8-10] which is necessary because the SRM model of Atashpaz-Gargari et al[5] is complex and does not have an analytical formulation of the likelihood. After calibration of the parameters, the ABC method is implemented via a Markov chain Monte Carlo (MCMC) procedure[13,15] to obtain a sample from the posterior distribution of the protein concentrations. Small MCMC sample sizes are sufficient to obtain a kernel-based implementation of the Optimal Bayesian Classifier (OBC).[12] Extensive experimental results examining the effect of various parameters demonstrate that the proposed method outperforms classical methods such as linear discriminant analysis (LDA) and 3NN[9], when sample size is very small, dimensionality is large, the data are noisy, or a combination of these

1.1 Organisation

The organization of the article is as follows. **Chapter 2:** SRM based MS model surveys the SRM-MS model. **Chapter 3:** ABC MCMC classification algorithm explains in detail the ABC rejection algorithm and the approximate Bayesian computation Markov chain Monte Carlo (ABC-MCMC) classifier. Section **Chapter 4:**Numerical experiments and results presents the numerical results. **Chapter 5:** Conclusions presents concluding remarks.

2. SRM-BASED MS MODEL*

In this article, we employ the model for the SRM pipeline proposed in the work by Atashpaz-Gargari et al [7].Next, we review briefly each of the main components of this model.

2.1 Protein Mixture Model

The protein mixture model concerns the true abundance of proteins in the SRM experiment. There are n samples in each class; for convenience, the two classes are labeled as 0 for control and 1 for treatment. There are N_{pro}^{a} proteins, N_{pro}^{c} of which are low-abundance candidates for biomarker validation. Protein identities are input as a FASTA file. As argued in previous works[1] , protein concentration can be modeled by a gamma distribution. Hence, the protein concentration is given by

$$\gamma_i \sim \begin{cases} \Gamma(k_c, \theta_c)\,, & i = 1, 2, 3, \ldots, N_{pro}^{c}\,, \\ \Gamma(k_a, \theta_a)\,, & i = N_{pro}^{c}+1, N_{pro}^{c}+2, \ldots, N_{pro}^{a}\,. \end{cases} \tag{2.1}$$

The variables k and θ are respectively shape and scale parameters. These are uniform random variables defined as $\mathrm{k}_c \sim$ Unif(k_c^{low},k_c^{high}), Unif(k_a^{low},k_a^{high}) and $\theta_c \sim$ Unif(θ_c^{low}, θ_c^{high}) , $\theta_a \sim$ Unif(θ_a^{low}, θ_a^{high}) respectively. The initial values of these variables, which are displayed in Table 1, reflect the dynamic range of protein abundance levels while taking into account that the candidate proteins are expressed at a much lower level than the background proteins. The initial values used here are consistent with values obtained experimentally in the work by Taniguchi et al as well as the hyperparameter values used in the work by Atashpaz-Gargari et al [7]. Furthermore, these initial values are modified based on the data, as part of the prior calibration process described in Algorithm 1.

*Reprinted with permission from *Bayesian Classification of Proteomics Biomarkers from Selected Reaction Monitoring Data using an Approximate Bayesian Computation-Markov Chain Monte Carlo Approach* by Kashyap Nagaraja and Ulisses Braga-Neto from Cancer Informatics journal Volume 17, pages 1:7, DOI:10.1177/1176935118786927, Copyright Date: May 24, 2018, Owners: Kashyap Nagaraja and Ulisses Braga-Neto

Parameter	Symbol	Value/Range
Instrument response factor	κ	5
Noise severity	α, β	0.03,3.6
Shape(Gamma distribution)	k_a, k_c	Unif(1.6,2.4), Unif(4,6)
Scale(Gamma Distribution)	θ_a, θ_c	Unif(9e6,11e6), Unif(90,110)
Purification	η_i	10^{-6}
Coefficient of Variation	ϕ	Unif(0.3,0.5)
Fold change	$\mathbf{f}$	Unif(1.5,1.6)
peptide efficiency factor	e_i	[0.1,1]

Table 2.1: SRM Instrument parameters

Proteins are divided into biomarker (differentially expressed) and nonbiomarker (not differentially expressed) proteins. We use fold change to quantify the difference:

$$f_l = \begin{cases} a_i\,, & \text{if the protein i is over expressed,} \\ \frac{1}{a_i}\,, & \text{if the protein is under expressed,} \\ 1, & \text{otherwise,} \end{cases} \tag{2.2}$$

for $l = 1, \ldots, N^a_{pro}$. The fold change parameter a_i is uniformly distributed in the interval $[1, h]$, for $h > 1$. The value of h used here is displayed in Table-1.

While the gamma distribution is chosen for mean protein concentrations, the variation of protein concentration is modeled by a multivariate gaussian vector. Accordingly, the concentration of protein l in class j is modeled as follows:

$$C^{pro}_{ij} \sim \begin{cases} N([\gamma_1, \gamma_2 \ldots, \gamma_{N^a_{pro}})], \Sigma)\,, & \text{for } j \in \text{class 0,} \\ N([\gamma_1 f_1, \gamma_2 f_2 \ldots, \gamma_{N^a_{pro}} f_{N^a_{pro}})], \Sigma)\,, & \text{for } j \in \text{class 1,} \end{cases} \tag{2.3}$$

for $l = 1, \ldots, N^a_{pro}$. Here we consider a diagonal covariance matrix $\Sigma = [\sigma^2_{lk}]_{N_{pro}XN_{pro}}$ so that the

protein concentrations are mutually independent or very weakly correlated (correlation between proteins can be included at the cost of adding more parameters to the model):

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_{22}^2 & 0 & \dots & 0 \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & \sigma_{N_{pro}^a}^2 \end{bmatrix}, \tag{2.4}$$

where

$$\sigma_{ij}^2 = \begin{cases} \sigma_{ii}^2, & \text{if } i = j \text{ and } i, j = 1, \dots, N_{pro}^a \\ 0, & \text{otherwise,} \end{cases} \tag{2.5}$$

and

$$\sigma_{ii}^2 = \phi * \gamma_{ii}^2, \qquad i = 1, \dots, N_{pro}^a \,. \tag{2.6}$$

The coefficient of variation ϕ has the initial value displayed in Table 1, which is the same as the one used in the work by Banerjee and Braga-Neto[1]. This value is modified based on the data, as part of the prior calibration process described in Algorithm 1. To model the purification process usually performed as part of the SRM-MS protocol, we select a set Gp of high-abundance proteins to be removed (in fact, attenuated) from the protein mixture:

$$C_{ij}^{\widehat{pro}} = \begin{cases} \eta_i C_{ij}^{pro}, & \text{for } i \in G_p, \\ C_{ij}^{pro}, & \text{otherwise.} \end{cases} \tag{2.7}$$

The value for η_i corresponds to the efficiency of the purification process and should be very small. See Table 1 for the value used in our simulation.

2.2 Peptide mixture model

In SRM-MS, tryptic digestion of proteins is carried out to generate small-mass peptides. Let Ω_i be the set of all the proteins which contain the i-th peptide.

$$C_{ij}^{pep} = \sum_{k \in \Omega_i} C_{kj}^{\widehat{pro}} \qquad \text{i=[1,2,...N}_c^{pp}\text{], j } \in [0,1] \tag{2.8}$$

The readout abundance μ_{ij} of the peptide can be modeled as

$$\mu_{ij} = C_{ij}^{pep} e_i \kappa \tag{2.9}$$

Here e_i represents the peptide efficiency factor and κ represents the SRM-MS response factor.

However, the true peptide abundance is different from its readout value due to the noise:

$$\nu_{ij} = \epsilon_{ij} + \lambda_{ij} \qquad \text{i=[1,2,...N}_c^{pp}\text{], j } \in [0,1] \tag{2.10}$$

where ϵ_{ij} is the additive gaussian noise which has a quadratic dependence on μ_{ij} as given below:

$$\epsilon_{ij} \sim N(0, \alpha\mu_{ij}^2 + \beta\mu_{ij}) \qquad \text{i=[1,2,...} N_c^{pp}\text{], j } \in [0,1] \tag{2.11}$$

where λ_{ij} is the additive exponential noise introduced due to transition effects.

$$\lambda_{ij} \sim exp(\mu_{tran}\mu_{ij}) \tag{2.12}$$

where μ_{tran} is a fixed constant.

The next step is called protein abundance roll-up. This is the process of obtaining the abundances of the parent proteins from the abundances and related characteristisc of their child peptides, detected during the MS1 process. To obtain the identities of the parent proteins, a second round of MS, called MS/MS, is often used and available databases of identities are searched. Here, we assume that the data from the rolled up abundances can be obtained and the readout of protein l in

sample j is given by

$$x_{lj} = \frac{1}{\kappa \eta_l} \sum_{i \in N_l} \nu_{lj} \qquad \text{l=[1,2,...} N_{pro}\text{], j} \in [0,1] \tag{2.13}$$

Where κ is the instrument response factor, N_l is the set of proteins present in peptide l and η_l is the number of peptides in set N_l. The data x_{lj} obtained in equation 2.13 are used for classification.

3. ABC-MCMC CLASSIFICATION ALGORITHM*

As described in the introduction section, the algorithm mainly has 3 steps: prior calibration via ABC rejection sampling, posterior sampling using an ABC-MCMC algorithm, and classification using a kernel-based method. We describe each of these steps below

3.1 Prior calibration via ABC rejection sampling

Once the protein abundances as described in equation (2.7) are obtained, the total number of proteins N pro a is reduced via a feature selection algorithm. As per the equations in the previous section, the protein abundance profiles are a function of the following:

- baseline parameters $\gamma = [\gamma_1, \gamma_2 \ldots ..\gamma_d]$
- Prior hyperparameters: $k_a, k_c, \theta_a, \theta_c, \phi, f$
- Instrument parameters: $\kappa, \alpha, \beta, e_i$

Prior calibration via ABC rejection sampling is as described in Algorithm 1. Monte Carlo integrations are performed to obtain a set of parameters and only some of them are kept and rest are rejected via comparing with a threshold.In this algorithm ϵ is the error tolerance. This has to be chosen optimally so that it should not be too high for bad samples to be accepted or it should not be very small that all the samples are accepted, i.e. $P(\|\mathbf{T}(S_0^{(t)}), \mathbf{T}(S_0)\| < \epsilon) \approx 0$ .Once the optimal parameters are obtained, the fold change vector is calculated by the following sample mean estimate:

$$f_{l,cal} = \frac{T_l(S_1)}{T_l(S_0)}, \qquad \text{l=0,1,2...,d} \tag{3.1}$$

where T_l denotes the l-th sample mean for the selected protein only.

*Reprinted with permission from *Bayesian Classification of Proteomics Biomarkers from Selected Reaction Monitoring Data using an Approximate Bayesian Computation-Markov Chain Monte Carlo Approach* by Kashyap Nagaraja and Ulisses Braga-Neto from Cancer Informatics journal Volume 17, pages 1:7, DOI:10.1177/1176935118786927, Copyright Date: May 24, 2018, Owners: Kashyap Nagaraja and Ulisses Braga-Neto

Algorithm 1 Prior calibration of $k_c, k_a, \theta_c, \theta_a, \phi$ using ABC Rejection sampling

1. Generate M_{cal} quintuplets of parameters of $k_c, k_a, \theta_c, \theta_a, \phi$ such that,

$$\begin{aligned} k_a^{(t)} &\sim Unif(k_a^{low}, k_a^{high}) \\ k_c^{(t)} &\sim Unif(k_c^{low}, k_c^{high}) \\ \theta_a^{(t)} &\sim Unif(\theta_a^{low}, \theta_a^{high}) \\ \theta_c^{(t)} &\sim Unif(\theta_c^{low}, \theta_c^{high}), \\ \phi^{(t)} &\sim Unif(\phi^{low}, \phi^{high}) \\ & \textbf{for } \; t = 1, 2, \ldots..M_{cal} \end{aligned}$$

2. Now simulate a control sample set $S_0^{(t)}$ of size n for each quintuplet of parameters **for** t=1,2....M_{cal}

3. Accept the quintuplet $(k_a^{(t)}, k_c^{(t)}, \theta_a^{(t)}, \theta_c^{(t)}, \phi_a^{(t)})$ if $\|\mathbf{T}(S_0^{(t)}) - \mathbf{T}(S_0)\| < \epsilon$, **for** $t = 1, 2 \ldots . M_{cal}$. Here $\|\|$ denotes the Euclidean norm and **T** denotes vector sample mean.

4. Let B=$[(k^1, \theta^1, \phi^1), \ldots., (k^{n_a}, \theta^{n_a}, \phi^{n_a})]$ be the set of accepted triplets.

5. The calibrated k can be approximated as follows:

$$k_a^{cal} = \int_{k_a^{low}}^{k_a^{high}} k_p(k_a|S_n)dk = \frac{1}{n_a} \sum_{a=1}^{n_a} k_a^{cal}$$

6. Similarly other four parameters are also calculated

3.2 ABC-MCMC Posterior Sampling

ABC-MCMC sampling is as described in Algorithm 2. Vector γ=$(\gamma_1, \gamma_2 \ldots . \gamma_d)$ is sampled from $p(\gamma|S_n) \propto p(S_n|\gamma)p(\gamma)$. After a burn-in period for the Markov chain of t_s, the next M samples from t_s to $t_s + M$ are considered as the generated data. Proper selection of the thresholds in step 4 of Algorithm 2 plays a very important role in the performance of the ABC-MCMC algorithm.

Algorithm 2 Obtain the posterior samples of γ using ABC-MCMC algorithm

1. Generate the mean vector $\gamma^{(0)} = (\gamma_0, \gamma_1, \ldots \gamma_d)$ from the Γ distribution with optimal parameters generated in algorithm 1.

For t=0,1,$\ldots t_s, t_{s+1}, \ldots . t_s + M$ where t_s is the burn-in period do:

2. Generate $\gamma^{(t+1)}$= ***ColMeans***$(S_0^{(t)})$ where ***ColMeans*** is a function which calculates mean feature(protein) wise.

3. Simulate the control and treatment samples S_0^{t+1} and S_1^{t+1} each of size using $\gamma^{(t+1)}$ and $\gamma^{(t+1)}.f_{cal}$ respectively.

4. Let

$$\text{q=} \begin{cases} 1 & \|\mathbf{T}(S_0^{(t+1)}) - \mathbf{T}(S_0)\| < \epsilon_0 \text{ and } \|\mathbf{T}(S_1^{(t+1)}) - \mathbf{T}(S_1)\| < \epsilon_1 \\ 0 & otherwise \end{cases}$$

5. If q=1, accept $\gamma^{(t+1)}$ else $\gamma^{(t+1)} = \gamma^{(t)}$

3.3 Kernel-Based Classification

We employ the kernel-based scheme proposed in the work by Banerjee and Braga-Neto,[3] which is itself based on the OBC in Dalton and Dougherty.[12] One of the issues with kernel based classification is choosing the right value of the kernel bandwidth parameter. If the value of the bandwidth parameter chosen is high, then it leads to oversmoothing and thus hiding many details in the data distribution. However, a small value for the bandwidth parameter leads to undersmoothing and thus many spurious noisy elements in the data are not eliminated. To address this, we employ an ensemble method, where different classifiers with different bandwidth parameters are obtained and then majority vote is used for classification. The classification algorithm is described in detail in Algorithm 3.

Algorithm 3 Using the ABC-MCMC based posterior samples for classification.

1. Choose a set of kernel bandwidth parameters $h = (h_1, h_2, \ldots, h_f)$ where f is the number of bandwidth values taken.

2. Choose the number of γ samples from markov chain to be used in the kernel classifier. Say we select **q** samples from the posterior. It is advisable to choose the samples from the end. For example in this case $t_s + M - q$ to $t_s + M$.

3. Choose a suitable kernel **K** for the analysis. In this paper we have chosen a zero mean unit variance gaussian kernel.

4.**For** a given test point **x** do:

. Declare a result vector res_vec=zeros[length(h)]

. **For** i in $h_1, h_2, \ldots, h_f$ do:

. **if** $(c \sum_{t=t_s+M-q}^{t_s+M} \sum_{j=1}^{n} \mathbf{K}(\frac{x-x^{(j)}}{h_i}) \geq (1-c) \sum_{t=t_s+M-q}^{t_s+M} \sum_{j=n+1}^{2n} \mathbf{K}(\frac{x-x^{(j)}}{h_i}))$

. res_vec[i]=1

. **else**

. res_vec[i]=0

5. The kernel based classifier is now given by,

$$\Psi\mathbf{(x)} = \begin{cases} 1 & \text{if } \mathbf{sum}(\text{res\_vec}) \geq \frac{f+1}{2}. \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

4. NUMERICAL EXPERIMENTS AND RESULTS*

In this section we demonstrate the application of proposed ABC-MCMC classification algorithm for SRM data, using a synthetic dataset generated from a subset of the human proteome. We selected a list of proteins from the Drugbank and applied tryptic digestion of proteins using the OpenMS software [1]. Since our interest is in small sample sizes we chose simple classification rules, which are known to perform well with small samples, for comparison: Linear discriminant analysis (LDA) and K-Nearest neighbor (KNN) with K=3.

Synthetic SRM-MS data were generated by the model described in section-2, using the parameters in Table-1. Synthetic sample data for prior calibration were generated using the midpoint of the intervals specified in Table-1. For example, since $\phi \sim$ Unif(0.3,0.5), we take 0.4 as the initial value.

For the MCMC procedure, we consider 10000 samples from the posterior distribution of γ. A burn-in stage of around 3000 iterations is considered. The value of prior probability was taken to be 0.5 (equally-likely classes). Kernel density estimation is based on 15 MCMC samples of γ, i.e **q=15** in Algorithm 3 (increasing this number did not show any significant difference in the results). From the initial number of 350 proteins, a t-test is applied to select the top 10-15 proteins. The t-test can select the protein features erroneously as opposed to more sophisticated feature selection methods, which makes the experiment more realistic.

We consider sample sizes $n = 10$ through $n = 40$ per class, and select the number of features to be $d = 3, 5, 8, 10$. A total of 6 runs of the experiment are conducted for each combination of classification rule, sample size, and dimensionality, and the average error for each case is obtained via a synthetic test dataset of 100 sample points

*Reprinted with permission from *Bayesian Classification of Proteomics Biomarkers from Selected Reaction Monitoring Data using an Approximate Bayesian Computation-Markov Chain Monte Carlo Approach* by Kashyap Nagaraja and Ulisses Braga-Neto from Cancer Informatics journal Volume 17, pages 1:7, DOI:10.1177/1176935118786927, Copyright Date: May 24, 2018, Owners: Kashyap Nagaraja and Ulisses Braga-Neto

4.1 Effect of Sample size

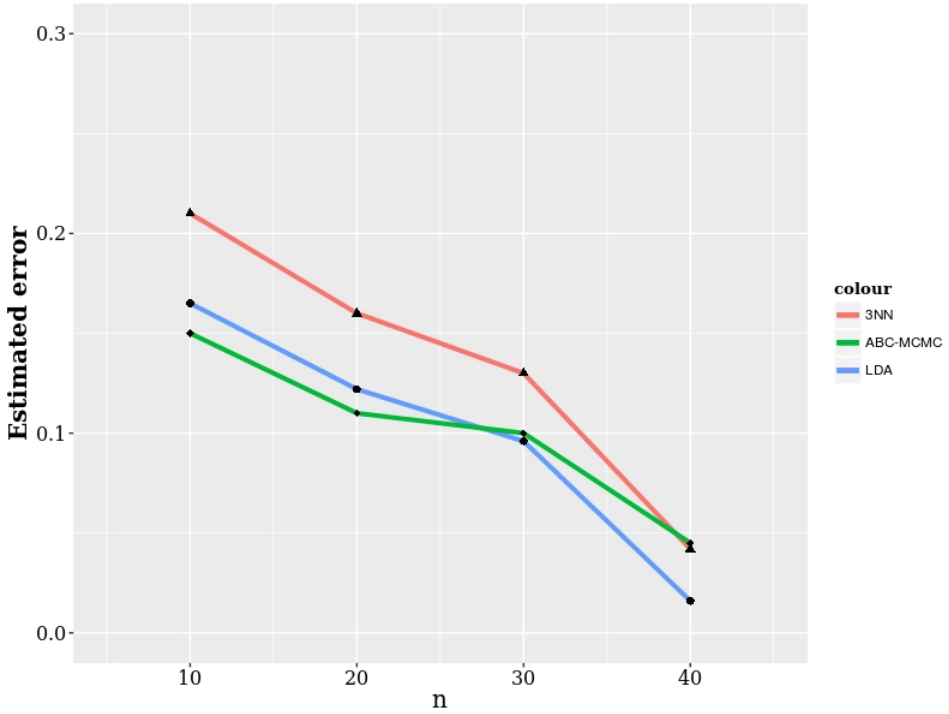

Figure 4.1: Variation of estimated error rate by number of samples.

Figure 1 displays the average error rates for the different classification rules. The number of proteins selected is fixed at $d = 10$. With the increase in sample size we see that the total error decreases for all classification rules. An important observation is that at small sample sizes, the performance of ABC-MCMC is best, confirming the general principle of good small-sample performance by Bayesian methods.

4.2 Effect of Dimensionality

The average error rates of the various classification rules against dimensionality, ie, number of selected proteins, are displayed in Figure 2, for fixed sample size n = 10 per class. We can observe a very strong peaking phenomenon16: as the number of selected proteins increases, the average classification error rates tend to go down at first, but then increase sharply, due to the small sample

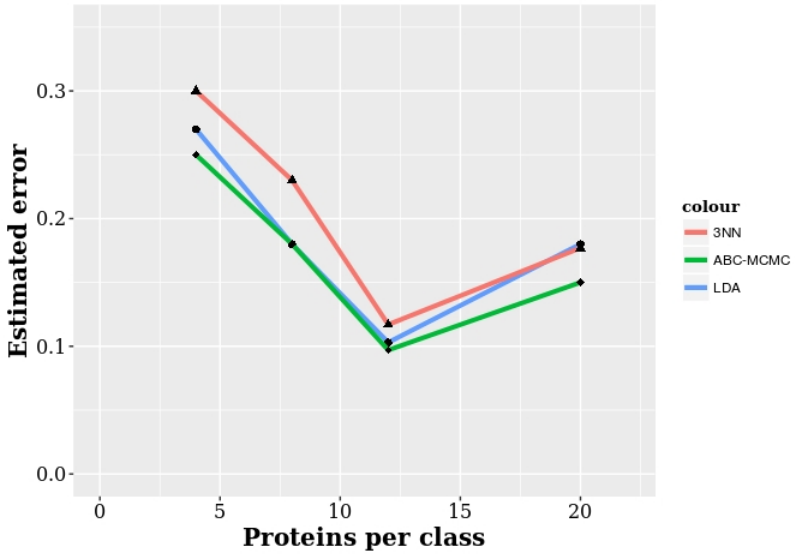

Figure 4.2: Variation of estimated error rate by the values of Dimensionality .

size, ie, small ratio between number of points over the dimensionality. One can observe that the ABC-MCMC classification rule is the most accurate one when d is large, which is in agreement with the fact that Bayesian methods tend to outperform competing techniques under small ratios of sample size to dimensionality.

4.3 Effect of Variability

Here, we keep the sample size at n = 10 and the number of features at d = 8 to investigate the impact on the classification of error rate of an increasing variability of the true protein concentration values. In Figure 3, one can observe that the performance of all classification rules degrades with increasing values of the coefficient of variation ϕ ; however, the performance of the ABC-MCMC algorithm is uniformly better than the others due to the small sample size n = 10 .

4.4 Effect of Peptide Efficiency

Finally, we investigate the impact on the classification accuracy of varying the peptide efficiency. The peptide efficiency factor a controls how many ions can be detected for a given peptide. Increasing this parameter uniformly increases efficiency for all peptides, which corresponds to a more accurate SRM-MS experiment. Indeed, one can observe in Figure 4 that classification

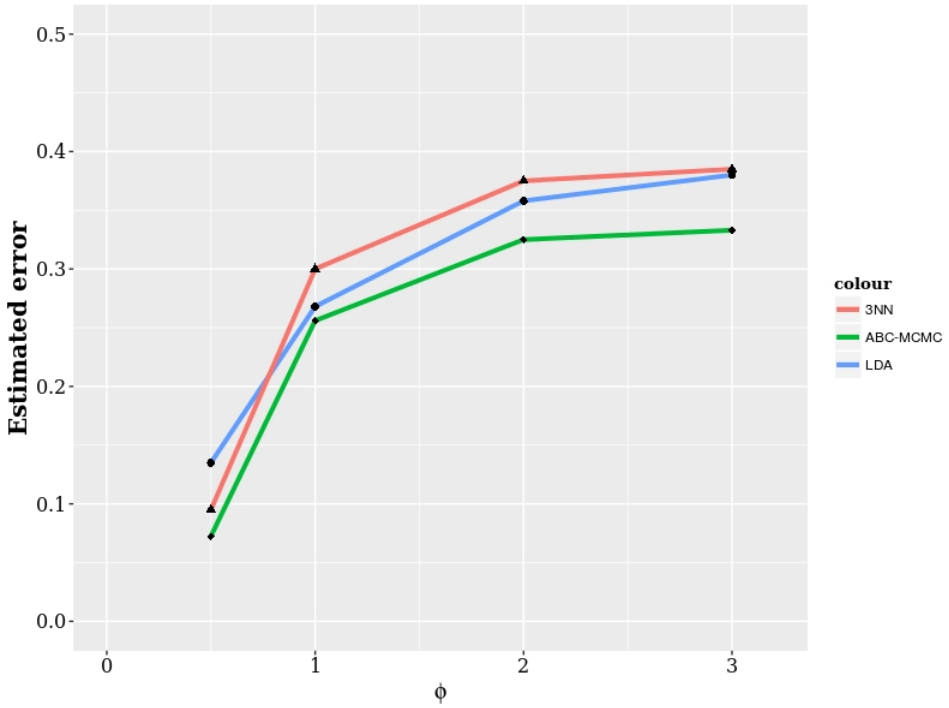

Figure 4.3: Variation of estimated error rate by the values of ϕ .

accuracy tends to increase with increasing peptide efficiency. One can also observe that the ABC-MCMC classification rule displays the smallest error rates among the competing methods at low peptide efficiency, ie, in a more noisy experiment.

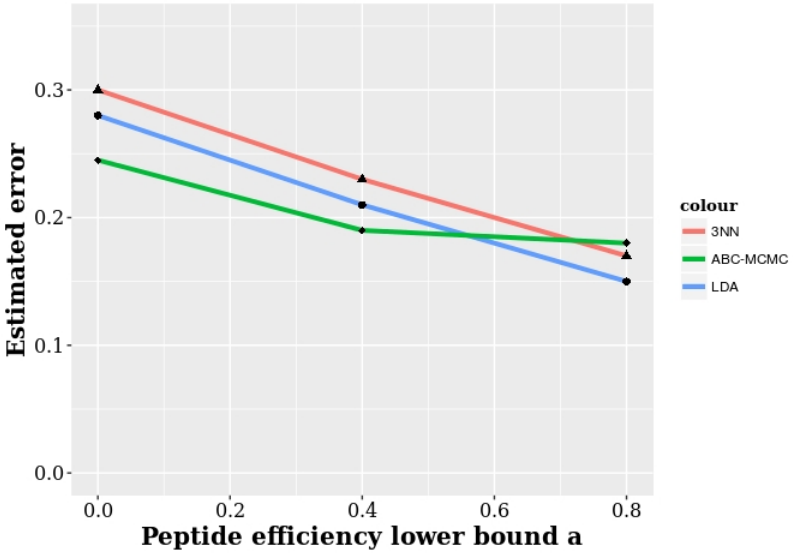

Figure 4.4: Variation of estimated error rate by the values of peptide efficiency .

5. CONCLUSIONS*

We have proposed a Bayesian approach for classifying SRM data with the goal of facilitating biomarker development. This method is a combination of ABC and MCMC. We can see that for small sample sizes, large dimensionality, or noisy data, the performance of the proposed Bayesian classifier is superior to that of other approaches. Our results are based on a subset of the human proteome selected from the Drugbank, which are submitted to tryptic digestion in silico. In addition, the prior hyperparameters are calibrated using the available data. This makes the the approach realistic and broadly applicable. Because we are studying the effects of the various parameters of the SRM pipeline on the classification error, there is a need to use synthetic data from a generative model. The results are, however, expected to be reproducible on clinical SRM data.

*Reprinted with permission from *Bayesian Classification of Proteomics Biomarkers from Selected Reaction Monitoring Data using an Approximate Bayesian Computation-Markov Chain Monte Carlo Approach* by Kashyap Nagaraja and Ulisses Braga-Neto from Cancer Informatics journal Volume 17, pages 1:7, DOI:10.1177/1176935118786927, Copyright Date: May 24, 2018, Owners: Kashyap Nagaraja and Ulisses Braga-Neto

REFERENCES

[1] U. Banerjee and U. Braga-Neto, "Bayesian ABC-MCMC classification of liquid-chromatography mass spectrometry data," *Cancer Informatics*, vol. Suppl. 5, pp. 175–182, 2017.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. IT-14, no. 1, pp. 55–63, 1968.

[3] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher, "Openms an open-source software framework for mass spectrometry," *BMC Bioinformatics*, vol. 9, p. 163, 2008.

[4] N.Rifai, M.Gillette, and S.Carr, "Protein biomarker discovery and validation: the long anuncertain path to clinical utility," *Nature Biotechnology*, vol. 24, pp. 971–983, 2006.

[5] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[6] R. Httenhain, J. Malmstrm, P. Picotti, and R. Aebersold, "Perspectives of targeted mass spectrometry for protein biomarker verification," *Curr. Opin. Chem. Biol,Nature*, vol. 13, pp. 518–525, 2009.

[7] E. Atashpaz-Gargari, U. Braga-Neto, and E. Dougherty, "Modelling and systematic analysis of biomarker validation using selected reaction monitoring," *Eurasip journal on Bioinformatics and Systems Biology*, 2014.

[8] B. Turner and I. V. Zandt, "A tutorial on approximate Bayesian computation," *Journal of mathematical Psychology*, vol. 56, pp. 69–85, April 2012.

[9] A. Webb, "Statistical pattern recognition," *New York: John Wiley & Sons 2nd ed*, 2002.

[10] K. Csillery, M. Blum, O. Gaggiotti, and O. François, "Approximate bayesian computation (ABC) in practice," *Trends in Ecology and Evolution*, vol. 30, no. 10, 2003.

[11] S. Sisson and Y. Fan, "Likelihood-free Markov chain monte carlo," *Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G. Jones, and X.-L. Meng eds.), Chapman and Hall/CRC Press*, 2010.

[12] L. Dalton and E. Dougherty, "Optimal classifiers with minimum expected error within a bayesian framework part i: Discrete and gaussian models," *Pattern Recognition*, vol. 46, no. 5, p. 1301âĂŞ1314, 2013.

[13] D. Wegmann, C. Leuenberger, and L. Excoffier, "Efficient approximate bayesian computation coupled with Markov chain monte carlo without likelihood," *Genetics*, vol. 182, pp. 1207–1218, August 2009.

[14] X. Ye, J. Blonder, and T. Veenstra, "Targeted proteomics for validation of biomarkers in clinical samples," *Brief. Funct. Genomics Proteomics*, vol. 8(2), pp. 126–135, 2009.

[15] C. J. Geyer, "Practical Markov chain monte carlo," *Statistical Science*, vol. 7, no. 4, pp. 473–483, 1992.